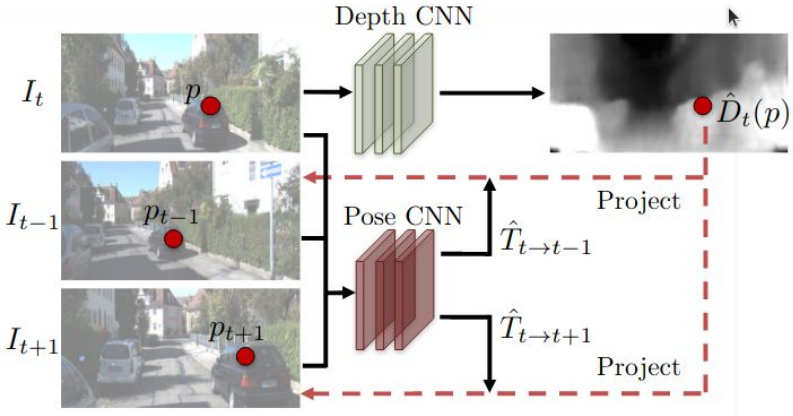
CMSC 733: Project 4
SfM Learner

Team Members:

Nishad Kulkarni - 117555431
Saurabh Palande - 118133959

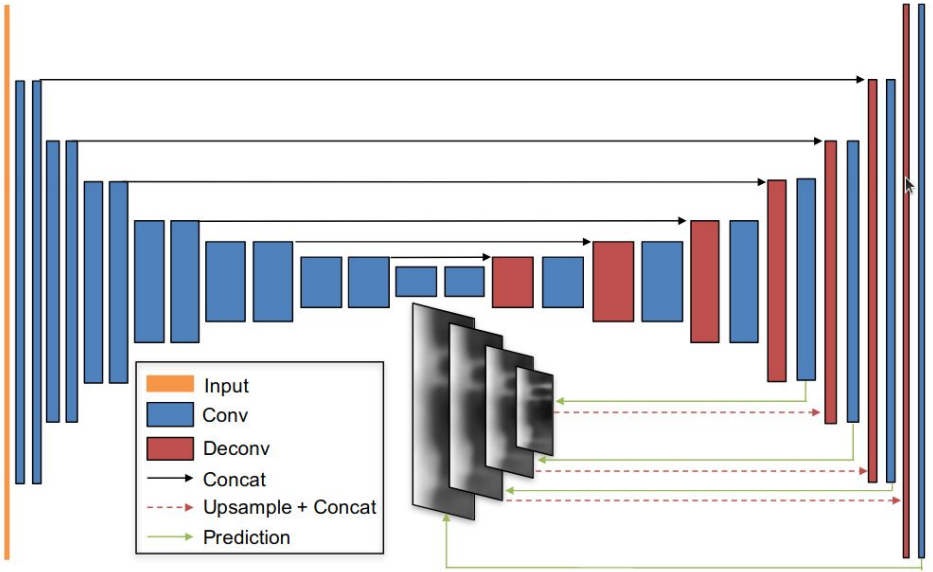
SfM learner - Overview

1. It is an end-to-end approach which maps directly from input pixels to an estimate of ego-motion (parameterized as 6-DoF transformation matrices)
2. The depth network takes only the target view as input, and outputs a per - pixel depth map D̂ t.
3. The pose network takes both the target view and the nearby/source views as input and outputs the relative camera poses



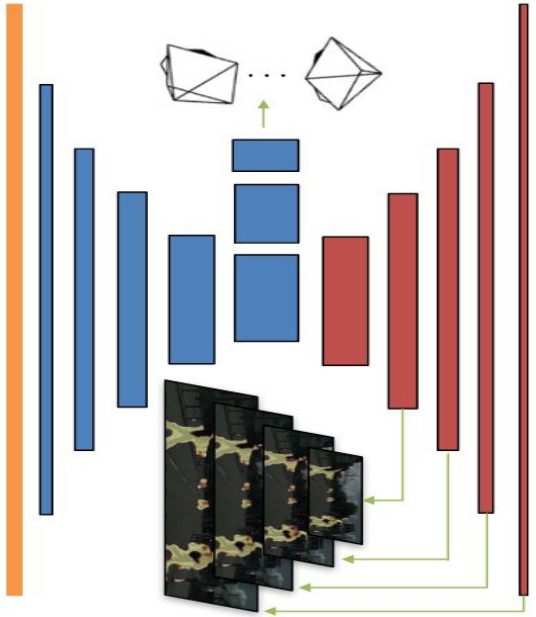
Depth Estimator

1. It uses the DispNet architecture that is mainly based on an encoder-decoder design with skip connections and multi-scale side predictions.
2. All conv layers are followed by ReLU activation except for the prediction layers



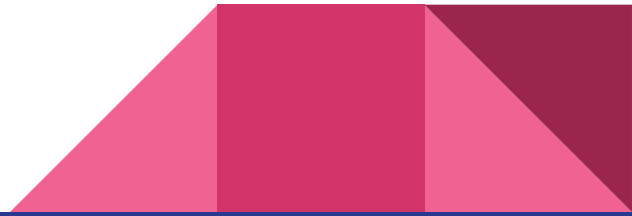
Pose/Explainability Network

1. Pose network consists of 7 stride-2 convolutions followed by a 1 * 1 convolution with 6 ∗ (N - 1) output channels.
2. Finally, global average pooling is applied to aggregate predictions at all spatial locations.
3. The explainability prediction network shares the first 5 feature encoding layers with the pose network, followed by 5 deconvolution layers with multi-scale side predictions.



Assumptions - SfM Learner

1. The scene is static without moving objects.
2. There is no occlusion/disocclusion between the target view and the source views.
3. The surface is Lambertian so that the photo-consistency error is meaningful



Network modification - 1

Using Structural Similarity loss

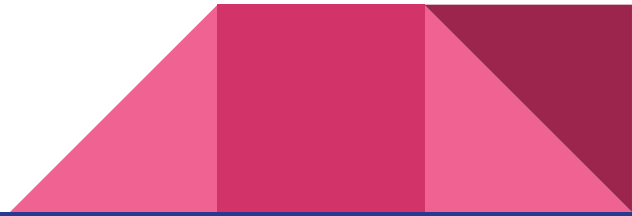
1. SfM-Learner uses photometric loss for training it's network.
2. It makes certain assumptions such as scenes need to have a constant brightness and luminosity.
3. To solve this problem, Structural similarity Index (SSIM) is used.
4. SSIM provides a robust metric for measuring perpetual differences between two images by considering the 3 factors of luminance, contrast and structure.

$$L_{ssim} = \sum_{s} \frac{1 - SSIM(I_t, I_s)}{2}$$

Network Modification -2

Using Multiple views for Depth Estimation

1. The SfM-Learner only takes a single target view as input to compute depth map.
2. By using both target and source views as input it provides a sharper depth map image and the network also tends to converge faster.
3. The multiple views leverages the relationship between pixels over multiple views to calculate depth and hence provides a better depth map.



Network Modification -3

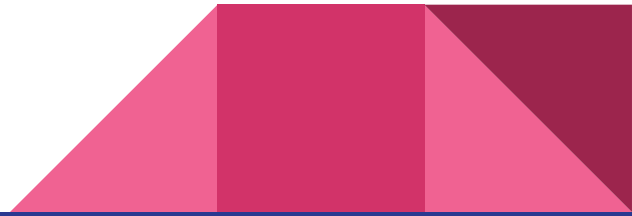
Using Adaptive Learning Rate

1. Learning rate of α = 0.0002 is used in the original implementation.
2. We implemented an adaptive learning rate, which starts with a high value of 0.0002 but gradually reduces at every step of 10,000 iterations

Network Modification -4

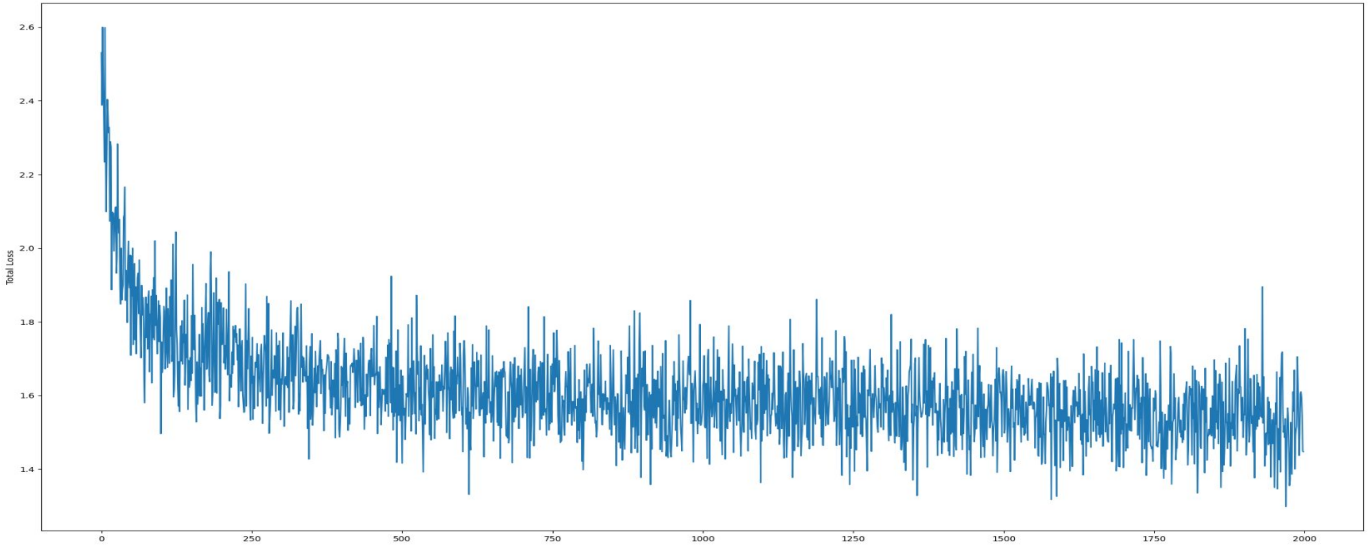
Data Augmentation

1. SfM-Learner already includes some data augmentation by randomly scaling and cropping the input data.
2. We added random shifting of the gamma values and randomly changing brightness and color of the images.



Results:

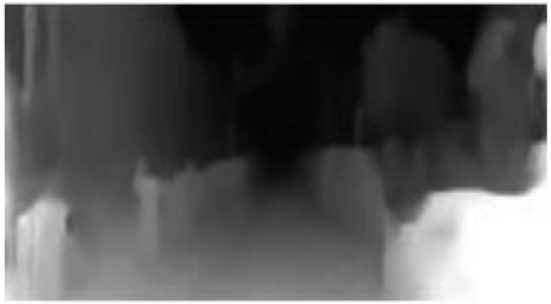
Loss Graph:



Depth map

Original Image

SfM learner -
Depth Map



Our model -
Depth map